Hector²

Spécialités

Les traitements spéciaux regroupent dans Hector des fonctionnalités qui, soit sont d'un usage trop rare ou trop complexe pour qu'on les impose à l'usager courant, soient relèvent d'approches statistiques originales, expérimentales voire hétérodoxes.

En effet, Hector est un logiciel destiné à couvrir un vaste ensemble de besoins, depuis ceux de la didactique des statistiques en premier cycle universitaire jusqu'à ceux qu'expriment les doctorants et les équipes de recherche. Pour autant, Hector ne prétend nullement être exhaustif : il y a des logiciels pour cela dont les prix s'écartent de celui de Hector de plusieurs ordres de grandeur. En fait, au delà d'un noyau de fonctionnalités choisies de manière à composer une boîte à outils raisonnablement complète, les éléments ajoutés à Hector résultent des propres besoins de recherche de l'auteur et des demandes émanant de la tribu de chercheurs utilisateurs de Hector.

C'est dire que les fonctionnalités offertes par la page SPECIALITES sont à utiliser par chacun à son gré, mais sous sa propre responsabilité, et que l'ergonomie peut en être un tantinet plus rugueuse que dans la page TRAITEMENTS. Il ne s'agit pas d'un véhicule de série, mais de prototypes, et la forme du rétroviseur est à prendre ou à laisser.

Pour ces raisons, la page SPECIALITES est exclusivement accessible aux titulaires d'une licence professionnelle-recherche, qui savent qu'ils s'aventurent en ces lieux à leurs risques et périls ; souvent, ils l'ont bien voulu car ce sont eux qui ont demandé ces fonctionnalités.

Chaque traitement spécial ou famille de traitement spéciaux est regroupée dans un volet de la page SPECIALITES.

LE VOLET DICHOTOMIES

Origine

Cette fonctionnalité assez hétérodoxe, qui s'apparente à une forme primitive de data mining, n'est pas issue de la demande d'un usager du logiciel, mais relève d'une vieille idée de l'auteur : la « moulinette suprême », celle qui dit si les données ont, oui ou non quelque chose à dire, même si elles essaient de le cacher.

Un chercheur bien organisé en a rarement besoin, tant il est vrai qu'on va aux données armé d'hypothèses à vérifier. Toutefois, l'enseignant de statistique y a souvent recours, pour répondre à un étudiant qui se plaint que ses données « ne contiennent rien » que, mais si, regardez, là, cette variable, si vous la regroupez comme ceci et que vous la croisez avec cela … D'accord, c'est de la triche, mais pour le bon motif.

Principe de la dichotomie

Dichotomie est un mot d'origine grecque qui désigne l'action de couper en deux parties.

On en connaît sans doute l'usage le plus fréquent en statistiques, qui est la dichotomie sur la médiane, laquelle, appliquée à une variable numérique ou simplement ordinale, place une coupure de telle sorte que les effectifs des deux parties soient approximativement égaux.

Cependant, les coupures en parts inégales sont aussi des dichotomies. N'importe quelle sorte de variable peut faire l'objet d'une dichotomie, à condition de respecter les caractéristiques de son type.

Dichotomie sur une variable numérique ou calendaire

Une variable numérique peut être découpée n'importe où, pourvu que ce soit *entre* deux valeurs existantes. En effet, même si on se représente les nombres, notamment décimaux, comme des entités continues, dans un ordinateur ils se ramènent toujours d'une manière ou d'une autre à des nombres entiers. Dans Hector, qui utilise le principe dit de « virgule fixe », le nombre de chiffres après la virgule est une caractéristique d'une variable numérique. Ainsi, avec une numérique à un chiffre après la virgule, il est possible de couper entre 14,2 et 14,3, mais non entre 14,19 et 14,2 , puisqu'il y faudrait deux décimales.

Le nombre de coupures distinctes qu'on peut opérer sur une variable numérique est le nombre de valeurs distinctes que prend cette variable, diminué de 1. Ainsi une numérique qui aurait comme valeurs 1, 5 et 12 peut subir deux dichotomies distinctes : 1 / 5 12 et 1 5 / 12.

La dichotomie sur une variable calendaire obéit au même principe, à ceci près que les calendaires sont représentées dans le modèle sous-jacent comme des nombres entiers de jours (donc, pas de décimales).

Dichotomie sur une variable logique

Une seule dichotomie possible, puisqu'une telle variable est dichotomique par nature : Faux / Vrai.

Dichotomie sur une variable ordinale

Même principe que pour les numériques : autant de dichotomies que de valeurs distinctes, moins une. Ainsi une variable ayant pour valeurs 'pas du tout', 'un peu', 'beaucoup', 'passionnément' peut subir trois coupures : une juste avant chaque valeur à l'exception de la première.

Dichotomie sur une variable nominale

Les choses sont nettement différentes, parce que comme les valeurs ne sont pas ordonnées, il s'agit de combinaisons, ou plus exactement de partitions complémentaires d'un ensemble. Un exemple aidera à fixer les idées.

Dans le tableau ci-dessous, on affiche les différentes répartitions possibles des valeurs a, b, c d'une variable nominale : selon qu'une valeur offre 0 ou 1, elle est dans la première partie de la dichotomie ou dans la seconde :

c	b	a
~~0~~	~~0~~	~~0~~
0	0	1
0	1	0
0	1	1
~~1~~	~~0~~	~~0~~
~~1~~	~~0~~	~~1~~
~~1~~	~~1~~	~~0~~
~~1~~	~~1~~	~~1~~

A première vue, pour trois objets, 2^n répartitions, comme pour les parties d'un ensemble. Cependant certaines sont barrées : 0 0 0 parce que triviale et sans intérêt, tous les objets étant dans la même partie, d'autres, à partir du milieu du tableau, parce qu'elles ne sont que le miroir des précédentes, et s'en déduisent par un échange des deux parties. 1 1 1, enfin, est doublement barrée parce qu'elle émarge aux deux motifs.

On remarque que les 0 et les 1 forment aussi les nombres de 0 à 2^3, la partie utile allant de 1 à 2^2-1 : ces numéros serviront à identifier les dichotomies de nominales.

On généralise en notant que n objets engendrent 2^{n-1}-1 dichotomies. C'est une quantité qui croît rapidement avec le nombre d'objets ; il est donc conseillé d'éviter d'en abuser. D'ailleurs les nominales a plus de 30 valeurs distinctes ne sont pas prises en compte par Hector dans cette fonctionnalité.

Ce que fait le programme

Il considère deux listes de variables, et croise entre elles toutes les dichotomies issues de chacune de ces variables, comme si chacune de ces dichotomies était une variable binaire ou logique. On notera que, si les listes sont conséquentes, on se retrouve très vite avec des milliers de croisements.

Pour chacun des croisements de dichotomies, calcule une statistique φ, qui présente le double avantage, dans ce cas précis d'un tableau 2x2, d'être analogue à un coefficient de corrélation des numériques (r de Bravais-Pearson, défini de -1 à +1) et d'être associée au χ^2 des nominales par la relation $\chi^2 = N \varphi^2$.

Considérant que *N* est le nombre total de sujets, *a* le nombre de sujets appartenant à la deuxième partie de la première variable, *b* le nombre de sujets appartenant à la deuxième partie de la deuxième variable et *c* le nombre de sujets réunissant ces deux caractéristiques :

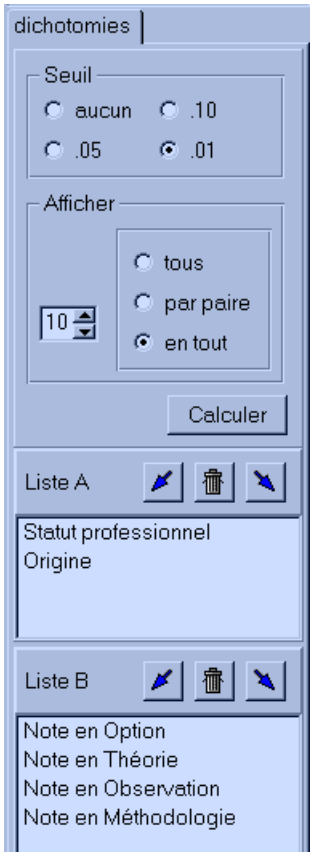
	Var B : 0	Var B : 1	Ensemble
Var A : 0			
Var A : 1		c	a
Ensemble		b	N

la formule de la statistique est alors :

$$\varphi = \frac{Nc - ab}{\sqrt{a(N-a)b(N-b)}}$$

Selon divers critères absolus (seuils de probabilité) ou relatifs (limitation du nombre de croisements de dichotomies affichés par paire de variable ou pour la liste totale), cette statistique φ permet de ne conserver que les résultats les plus intéressants, tant il est vrai que trop d'information tue l'information.

Les organes et la manoeuvre



De haut en bas :

Sélection du seuil

Les croisements de dichotomies n'atteignant pas le seuil de probabilité sélectionné ne sont pas affichés.

Limitation du nombre de résultats affichés

Si la sélection est [tous], aucune limitation. Attention, le nombre de dichotomies peut être très élevé. A la limite, le programme peut se bloquer à cause de la seule gestion de l'affichage des résultats.

Si la sélection est [par paire], seuls sont publiées les *n* meilleures dichotomies (celles qui obtiennent les plus grandes valeurs absolues pour φ) pour chaque paire de variables d'origine, *n* étant la valeur du compteur placé à gauche du sélecteur.

Si la sélection est [en tout], même principe de limitation aux meilleurs, mais il s'agit alors des n meilleurs sur l'ensemble des dichotomies demandées.

Le bouton [calculer]

Il lance les calculs, selon les options sélectionnées, et à condition bien sûr que des variables soient sélectionnées dans les listes A et B.

Les listes de variables

Pour le logiciel, la distinction entre la liste A et la liste B ne revêt aucune signification particulière, puisque le test φ est symétrique. Cependant, il peut être de bonne méthode de placer en liste A des variables plutôt indépendantes, susceptibles « d'expliquer », et en liste B des variables plutôt dépendantes, « à expliquer ».

Pour placer des variables dans ces listes, on les sélectionne dans la liste générale placée à droite comme partout ailleurs dans Hector, et on clique la flèche bleue tournée vers la liste. Pour enlever des variables d'une des listes, on les y sélectionne, et on clique la flèche bleue tournée vers l'extérieur. Avec le bouton poubelle, on vide la liste.

On peut aussi bien sélectionner des collections entières, dans la liste générale des collections. Si une variable figure à titre individuel, mais aussi dans une collection, ou dans plusieurs collections sélectionnées, elle n'est prise en compte qu'une seule fois.

Une même variable peut figurer dans les deux listes : d'ailleurs, si on demande le calcul alors qu'une seule liste est garnie, son contenu est préalablement recopiée dans l'autre. Cependant une même variable n'est jamais croisée avec elle-même, le résultat étant trivial et dépourvu d'intérêt.

La manœuvre

Elle consiste donc à installer un choix de variables dans les listes, à vérifier ou modifier les options choisies, et à cliquer le bouton [calculer]. Le résultat s'affiche dans la page de compte-rendu, que l'on peut aussi, comme d'habitude, envoyer à l'imprimante ou dans un document.

Lecture des résultats

Les résultats ont typiquement l'allure suivante :

Recherche des dichotomies efficaces

Variables de la liste A	coupure A	Variables de la liste B	coupure B	Phi
Origine dicho 2		Note en Méthodologie /14		+0,19 ***
Origine dicho 22		Note en Méthodologie /12		+0,18 ***
Origine dicho 32		Note en Option /14		+0,19 ***
Origine dicho 32		Note en Option /18		+0,16 ***
Origine dicho 32		Note en Théorie /7		-0,16 ***
Origine dicho 6		Note en Méthodologie /13		+0,19 ***
Origine dicho 6		Note en Méthodologie /14		+0,17 ***
Statut professionnel dicho 4		Note en Théorie /10		+0,18 ***
Statut professionnel dicho 4		Note en Théorie /12		+0,18 ***
Statut professionnel dicho 4		Note en Méthodologie /12		+0,17 ***

Cette première partie du compte-rendu fournit les informations dans l'ordre suivant :

- nom de la variable A
- critère de coupure sur la variable A
- nom de la variable B
- critère de coupure sur la variable B
- valeur du test φ
- éventuelles étoiles selon la significativité du test

Les critères de coupure peuvent être exposés de deux façons différentes, selon le type des variables. Pour les variables appartenant à un type muni d'un *ordre* (toutes sauf les nominales), la coupure est manifestée par le caractère '/', suivi de la plus basse valeur (le début) de la seconde

partie de la dichotomie. Pour les variables nominales, il est fourni un numéro de dichotomie, qui sera détaillé en dessous, puisqu'une même dichotomie peut apparaître plusieurs fois.

Le test φ prend ses valeurs entre -1 et +1, comme un coefficient de corrélation qu'il est. Selon la convention de Hector, 3 étoiles veulent dire « significatif au seuil de .01 », 2, « .05 », 1, « .10 », et rien, « non significatif ».

Dans l'exemple ci-dessus, les options choisies étaient « seuil à .01 » et « 10 résultats en tout ».

La seconde partie du compte-rendu détaille, si besoin est, les dichotomies sur nominales :

```
Détail des dichotomies sur nominales

_____________________________________
Origine : dicho n°2
   D.Lettre D.ScHum Soc.Form Tech.Sup Val.acq
   D.Sces
Origine : dicho n°6
   D.Lettre Soc.Form Tech.Sup Val.acq
   D.Sces D.ScHum
Origine : dicho n°22
   D.Lettre Soc.Form Val.acq
   D.Sces D.ScHum Tech.Sup
Origine : dicho n°32
   D.Lettre D.Sces D.ScHum Soc.Form Tech.Sup
   Val.acq
Statut professionnel : dicho n°4
   aut.prof enseignt
   étudiant
```

Pour chaque dichotomie, la première ligne énumère les valeurs membres de la première partie, la seconde ligne énumère les valeurs membres de la seconde partie.

Stratégies d'utilisation

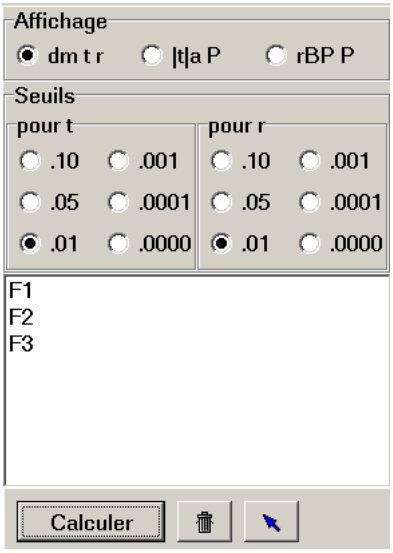
La recherche des dichotomies efficaces est un instrument rude mais puissant d'exploration des données. On peut se le représenter comme un chalut, un de ces filets géants que de gros bateaux de pêche traînent derrière eux. Comme le chalut, cette méthode peut ramener beaucoup de choses, et éventuellement un peu de tout. Il convient donc de s'en servir avec discernement, et d'affiner ensuite les résultats obtenus. Ces précautions énoncées, cet outil permet de savoir immédiatement dans quel secteur des données on va pouvoir trouver des relations significatives. Même si on ne les exploite pas sous la forme réduite de dichotomies, ces relations, quand elles existent, laissent nécessairement des traces sous cette forme. On peut continuer à exploiter les numériques comme telles, mais peut-être en les transformant en classes significatives : on peut par exemple ensuite fabriquer, comme ordinales, des classes d'âge non linéaires mais dont les frontières constitue des seuils significatifs au regard d'autres variables. Sans réduire toutes les nominales à des dichotomies, on peut parfois obtenir des regroupements en un plus petit nombre de classes, selon des découpages suggérés par les dichotomies. L'énorme avantage de la méthode est qu'elle s'applique à tous les types de variables simultanément, au lieu de devoir changer de test selon les types. Elle rend comparable ce qui ne l'était pas.

LE VOLET COLLECTIONS PARALLELES

Origine

Les collections parallèles sont issues des besoins de l'enquête GO de 2006 menée par la FUNOC de Charleroi (Belgique), avec Christou Verniers et Thérèse Levené. Il s'agissait initialement de comparer les représentations de stagiaires de la formation continue *avant* et *après* une formation longue : on leur avait donc soumis en deux fois la même série de valuateurs[1], et on comparait les résultats. Cette technique va être réemployée par le même organisme pour une étude intégrale de son public en 2008-2009, mais elle joue aussi un rôle important dans la démarche d'évaluation des formations de l'IUFM Nord-Pas de Calais menée de 2006 à 2008 par Bruno Perrault.

Organes



En haut, le sélecteur du mode d'affichage.

En dessous, les sélecteurs de seuils pour |t| et pour r.

En dessous, la liste des collections parallèles (ici elle est déjà garnie).

En dessous, le bouton [Calculer] qui fait ce que son nom indique, la poubelle pour vider la liste des collections parallèles, et la flèche bleue pour y installer une collection préalablement sélectionnée dans la liste des collections.

Démarche

On doit au préalable disposer d'une certain nombre (au moins deux) de collections *numériques* parallèles, ce qui signifie :

- qu'elles ont toutes le même nombre de variables
- que la première variable de chaque collection parallèle correspond à la première des autres, de même que la deuxième … les variables de même rang se *correspondent*.

[1] Un valuateur est l'association d'une proposition (une phrase) et d'une échelle à 8 cases, polarisée aux extrémités par un couple de verbalisation antagonistes (exemple : pas d'accord du tout / tout à fait d'accord). L'enquêté indique sa réponse par une croix dans une des huit cases, selon la nuance qu'il veut exprimer. (Définition A. Dubus in « Cahiers Theodile » n°1, 2000, Université de Lille 3).

Cette correspondance est de nature sémantique, le logiciel ne peut rien faire pour s'en assurer. C'est à l'usager de construire ses collections toutes sur le même modèle.

Dans le cas GO, la correspondance (cf. corpus go5_ad.cn) signifie qu'il s'agit des mêmes valuateurs présentés à trois moments successifs (mais ça pourrait signifier autre chose dans d'autres cas). Assez souvent, il s'agira de mesures avant/après, mais il peut s'agir aussi de jugements portés sur les mêmes objets par des classes d'observateurs différents.

On dispose donc des séries de collections F1 F2 F3 pour le rapport à la formation, E1 E2 E3 pour le rapport à l'emploi, O1 O2 O3 pour la représentation des obstacles.

Dans la liste des collections en bas à droite, on sélectionne F1, et on l'installe dans la liste des collections impliquées au moyen de la petite flèche bleue oblique.

On fait pareil avec F2 et F3. Si on se trompe sur l'ordre ou le choix, on vide la liste en cliquant sur la poubelle, et on recommence.

Affichage dm t r

Quand c'est OK, on clique le bouton calculer, et on obtient ceci :

```
Différences des moyennes, t : |t| s. à .01   ; r : rBP s. à .01
valeurs calculées parallèlement sur les variables des collections :
      1. F1
      2. F2
      3. F3
         1           2           3   1  ->  2    1  ->  3    2  ->  3
_____________________________________________________________________

   Curieux    Curieux2    Curieux3     0,59 t      0,68 t      0,16  r
 Isolement  Isolement2  Isolement3     0,24  r     0,34  r     0,08  r
RéflexProf  RéfleProf2  RéfleProf3     0,00        0,00       -0,03
PisteEmplo  PisteEmpl2  PisteEmpl3    -0,07        0,15        0,13
ObligaOnem  ObligOnem2  ObligOnem3    -0,35 tr    -0,34        0,03  r
     Utile      Utile2      Utile3    -0,31       -0,20        0,13
 ConseillE   ConseilE2   ConseilE3     0,02  r    -0,05  r     0,00  r
 Confiance  Confiance2  Confiance3     0,17  r     0,32  r     0,13  r
OrienForma  OrienForm2  OrienForm3    -0,36       -0,20        0,10
PasLeChoix  PasLeChoi2  PasLeChoi3     0,07        0,02        0,08  r
Occupation  Occupatio2  Occupatio3    -0,02  r     0,29  r     0,23  r
PossiEmplo  PossiEmpl2  PossiEmpl3    -0,10        0,00        0,08  r
    Argent     Argent2     Argent3     0,12  r     0,15  r     0,13  r
Rencontres  Rencontre2  Rencontre3     0,21  r     0,39 tr     0,18  r
  Suivisme   Suivisme2   Suivisme3    -0,14  r    -0,15  r     0,00  r
ConnaDeSoi  ConnDeSoi2  ConnDeSoi3     0,00  r     0,37  r     0,30  r
InfoEmpFor  InfEmpFor2  InfEmpFor3    -0,45       -0,32        0,15  r
AideRecEmp  AidRecEmp2  AidRecEmp3    -0,45 tr    -0,32  r     0,15  r
_____________________________________________________________________
```

Les colonnes 1, 2 et 3 listent les variables des trois collections parallèles.

Les trois colonnes suivantes traitent des comparaisons entre la première et la seconde collection (1->2), puis première troisième, puis deuxième troisième.

Si on a plus de trois colonnes, on aura en largeur toute la combinatoire des comparaisons des moments deux à deux. On préférera donc souvent fractionner le problème.

Dans ces colonnes, et parce qu'on a utilisé l'option par défaut [dm t r], on trouve :

dm : la différence des moyennes entre variables correspondantes

un t si la différence des moyennes des mesures est significative au seuil sélectionné pour les |t| de Student sur échantillons appareillés.

un r si le coefficient de corrélation de Bravais-Pearson est significatif au seuil pour le rBP

Ainsi, à la ligne ObligaOnem, colonne 1->2, on trouve :

`-0,35 tr`

Ce qui signifie :

que la différence de score moyen entre ObligOnem2 et ObligaOnem est de -0,35 (diminution)

que le |t| est significatif (au seuil sélectionné, ici .01)

que le rBP est significatif au seuil sélectionné (.01)

On note qu'on peut fixer un seuil de significativité différent pour le |t| de Student et pour le r de Bravais-Pearson, parce qu'il y a des situations où c'est utile de le faire, avec des mouvements de moyenne ténus mais de fortes corrélations, par exemple (ce pourrait être le cas ici).

Affichage |t|a P

Avec cette option, qui veut dire |t| de Student sur échantillons appareillés, avec le seuil P de probabilité, on obtient ceci :

```
|t| de Student sur échantillons appareillés, avec seuil de significativité
valeurs calculées parallèlement sur les variables des collections :
      1. F1
      2. F2
      3. F3
         1           2           3   1  ->  2    1  ->  3    2  ->  3
```

Curieux	Curieux2	Curieux3	2,96	.01	3,60	.001	1,18	
Isolement	Isolement2	Isolement3	1,40		2,01	.10	0,77	
RéflexProf	RéfleProf2	RéfleProf3	0,00		0,00		0,27	
PisteEmplo	PisteEmpl2	PisteEmpl3	0,50		0,97		0,93	
ObligaOnem	ObligOnem2	ObligOnem3	2,88	.01	1,93	.10	0,19	
Utile	Utile2	Utile3	2,47	.05	1,60		0,90	
ConseillE	ConseilE2	ConseilE3	0,15		0,30		0,00	
Confiance	Confiance2	Confiance3	1,19		2,17	.05	1,22	
OrienForma	OrienForm2	OrienForm3	2,15	.05	1,24		0,81	
PasLeChoix	PasLeChoi2	PasLeChoi3	0,38		0,14		0,60	
Occupation	Occupatio2	Occupatio3	0,16		1,82	.10	1,65	
PossiEmplo	PossiEmpl2	PossiEmpl3	0,57		0,00		0,68	
Argent	Argent2	Argent3	0,84		1,00		1,15	
Rencontres	Rencontre2	Rencontre3	1,46		3,39	.01	1,74	.10
Suivisme	Suivisme2	Suivisme3	1,18		1,23		0,00	
ConnaDeSoi	ConnDeSoi2	ConnDeSoi3	0,00		2,56	.05	2,08	.05
InfoEmpFor	InfEmpFor2	InfEmpFor3	0,96		0,65		1,43	
AideRecEmp	AidRecEmp2	AidRecEmp3	4,16	.001	2,48	.05	1,18	

Pour chaque comparaison de variables deux à deux, on a explicitement la valeur du |t| calculé et son seuil de significativité (ou rien s'il n'est pas significatif, fût-ce à .10). Pas de référence au rBP

Affichage rBP P

Avec cette option , qui signifie Coefficient de corrélation de Bravais-Pearson avec son seuil P de probabilité, on obtient ceci :

```
Coefficient de corrélation de Bravais-Pearson avec seuil de significativité
valeurs calculées parallèlement sur les variables des collections :
      1. F1
      2. F2
      3. F3
         1           2           3    1  ->  2     1  ->  3     2  ->  3
```

Curieux	Curieux2	Curieux3	0,30	.10	0,38	.05	0,59	.0001
Isolement	Isolement2	Isolement3	0,54	.001	0,57	.001	0,81	.0001
RéflexProf	RéfleProf2	RéfleProf3	0,27	.10	0,21		0,39	.05
PisteEmplo	PisteEmpl2	PisteEmpl3	0,32	.05	0,25		0,35	.05
ObligaOnem	ObligOnem2	ObligOnem3	0,69	.0001	0,35	.05	0,60	.0001
Utile	Utile2	Utile3	0,32	.05	0,17		0,09	
ConseillE	ConseilE2	ConseilE3	0,60	.0001	0,64	.0001	0,59	.0001
Confiance	Confiance2	Confiance3	0,51	.001	0,46	.01	0,66	.0001
OrienForma	OrienForm2	OrienForm3	0,29	.10	0,36	.05	0,34	.05
PasLeChoix	PasLeChoi2	PasLeChoi3	0,25		0,39	.05	0,64	.0001
Occupation	Occupatio2	Occupatio3	0,63	.0001	0,55	.001	0,66	.0001
PossiEmplo	PossiEmpl2	PossiEmpl3	0,25		0,19		0,63	.0001
Argent	Argent2	Argent3	0,68	.0001	0,68	.0001	0,82	.0001
Rencontres	Rencontre2	Rencontre3	0,53	.001	0,75	.0001	0,71	.0001
Suivisme	Suivisme2	Suivisme3	0,68	.0001	0,71	.0001	0,91	.0001
ConnaDeSoi	ConnDeSoi2	ConnDeSoi3	0,54	.001	0,61	.0001	0,52	.001
InfoEmpFor	InfEmpFor2	InfEmpFor3	0,18		0,18		0,64	.0001
AideRecEmp	AidRecEmp2	AidRecEmp3	0,67	.0001	0,57	.0001	0,61	.0001

On a les valeurs du coefficient de corrélation calculé, suivi du seuil de probabilité correspondant.

Interprétation

Si on ne s'intéresse qu'au progrès ou à la régression, le |t| de Student suffit. Si on s'intéresse plutôt à la cohérence des valeurs avant/après, le coefficient rBP suffit.

Toutefois la combinaison des deux peut amener des considérations intéressantes, selon le tableau suivant :

	rBP non significatif	rBP significatif
|t| non significatif	Pas de différence entre les moyennes, mais pas de cohérence non plus : mouvement brownien dans le nuage de points	Pas de différences entre les moyennes, mais une bonne cohérence des mesures : le nuage est stable et ne bouge pas
|t| significatif	Une différence significative (progrès ou régression), mais pas de cohérence : le nuage a bougé en se déformant.	Une différence significative (progrès ou régression), avec une forte cohérence : le nuage s'est déplacé sans se déformer.

Tout ceci à relativiser bien sûr selon les seuils que l'on choisit.

LE VOLET META-FORMULE

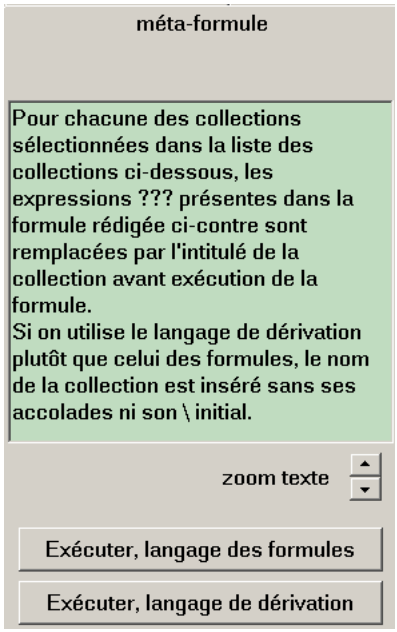
Origine

Cette fonctionnalité a été ajoutée en 2006 ou 2007 pour répondre aux besoins de Bruno Perrault, qui, dans le traitement des données de l'évaluation annuelle des formations de l'IUFM Nord-Pas-de-Calais, avait à effectuer sur de très nombreuses collections de nombreuses opérations répétitives. Il fallait donc trouver un moyen simple d'effectuer une tâche pénible et compliquée.

L'idée de la méta-formule repose sur l'idée de *méta-collection*, ou collection de collection.

Si on repart de la donnée de base dans un corpus : « Le sujet n°43 est une femme », autrement dit la valeur unique que prend une variable (Sexe) pour un sujet (43), un premier niveau d'abstraction est la variable, qui rassemble les valeurs de tous les sujets, et permet de les étudier d'un seul coup ou de les recouper avec d'autres variables. Un niveau d'organisation supérieur est la collection, qui rassemble plusieurs variables du même type, et permet de mener des opérations sur plusieurs variables simultanément. Et si l'on veut travailler sur plusieurs collections à la fois ? Alors il faut une entité de plus haut niveau, la méta-collection, regroupement de collections de variables du même type. Cette entité sera fugace, le temps du calcul auquel on la destine ; elle n'aura donc pas de nom, et ne sera pas stockée, contrairement aux collections.

Démarche


méta-formule
Pour chacune des collections sélectionnées dans la liste des collections ci-dessous, les expressions ??? présentes dans la formule rédigée ci-contre sont remplacées par l'intitulé de la collection avant exécution de la formule.
Si on utilise le langage de dérivation plutôt que celui des formules, le nom de la collection est inséré sans ses accolades ni son \ initial.
zoom texte
Exécuter, langage des formules
Exécuter, langage de dérivation

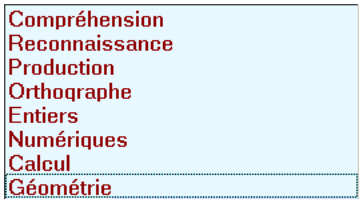
Le panneau vert explicite la règle du jeu. Le bouton d'édition permet d'agrandir les caractères[2] dans le panneau jaune à gauche, qui est un éditeur de texte[3].

Deux boutons d'exécution sont proposés, parce que deux langages sont possibles. Le langage des formules est l'officiel : il est propre à Hector[2] (2008) alors que le langage de dérivation est hérité de Nestor (2002). Obsolète, il est maintenu pour la compatibilité ascendante des corpus anciens.

[2] Certains d'entre nous n'ont plus vingt ans.

[3] Où l'on peut donc écrire, mais aussi copier, couper et coller de et vers d'autres éditeurs, comme un trtaitement de texte.

Pour disposer d'une méta-collection, il faut avoir sélectionné plusieurs collections[4] :



Les spécialistes de l'enseignement auront reconnu les *domaines* de compétence des tests nationaux CE2, version début XXIème siècle.

Chacune de ces collections rassemble plusieurs variables ; ainsi Compréhension rassemble les variables suivantes :

Il s'agit ici des différentes *compétences* du domaine Compréhension.

Un exemple simple

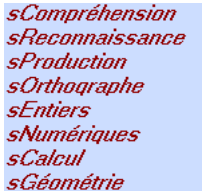
On tape dans l'éditeur le texte suivant :

```
# s???
: SOMME ??? ;
```

En gros, cela s'interprète ainsi[5] : pour chaque collection de la méta-collection, créer une variable numérique dont le nom sera le même que celui de la collection, avec un 's' ajouté devant, et dont la valeur sera pour chaque sujet la somme des valeurs des variables de la collection[6].

Ici il y a peu de risque d'erreur, car le cas est simple. Si on veut faire plus compliqué, il est conseillé de tester une maquette de la formule (avec une vraie collection au lieu des ???) dans la page des FORMULES, où est effectuée une vérification syntaxique absente ici.

Après exécution, on remarque que la liste des variables s'est allongée :



Chacune de ces variables, écrite en italique comme il convient aux formulées, contient la somme des valeurs de la collection correspondante.

Un exemple un peu moins simple

Sur une autre méta-collection provenant d'un autre corpus, on va tenter quelque chose d'un peu plus complexe, exploitant notamment le fait que le langage des formules permet d'agir au niveau de chaque variable, et pas seulement pour additionner des collections.

---

[4] Dans l'exemple on a sélectionné toutes les collections disponibles. Ca n'a rien d'obligatoire, on sélectionne ce que l'on veut du moment que ces collections sont du même type. On peut même sélectionner des collections non contiguës, avec Ctrl-clic, mais le plus souvent dans un corpus bien rangé, les collections apparentées sont contiguës, d'autant qu'on peut modifier leur ordre pour ce faire.

[5] Pour plus de rigueur, se reporter au manuel du Langage des Formules.

[6] Dans le langage de dérivation, cela s'écrirait plutôt :

```
var {s ???}
num % 0 100 0
som #\{???} fin
```

Certes, il y a une parenté, mais on peut constater qu'on a un peu gagné en transparence.

On a ci-dessus les collections moyennes des matières principales d'un grand nombre de lycéens aux trois moments cruciaux de l'année.

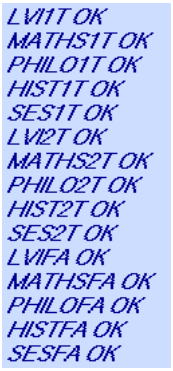
Chacune de ces collections rassemble plusieurs moyennes, ainsi les notes du premier trimestre :

On emploie la formule suivante :

```
POUR ??? + _OK \
£ NOUVELLE
: SOURCE >= 10 ;
```

qui peut s'interpréter ainsi : pour chacune des variables de chacune des collections de la méta-collection, , créer une nouvelle variable logique nommé comme la variable d'origine avec le suffixe « OK », et qui est vraie si la valeur de la variable d'origine est égale ou supérieure à 10.

Les variables suivantes sont créées :



Chacune d'elle signifie, par Vrai ou Faux, pour chaque élève, si celui-ci a réussi[7] ou non dans la discipline concernée ce semestre-là.

La compréhension fine du mécanisme passe par l'étude du manuel des Formules, mais on concèdera que le jeu peut valoir la chandelle.

[7] Selon le critère conventionnel de la réussite à partir de la note de 10.

LE VOLET LPE/CONSENSUS

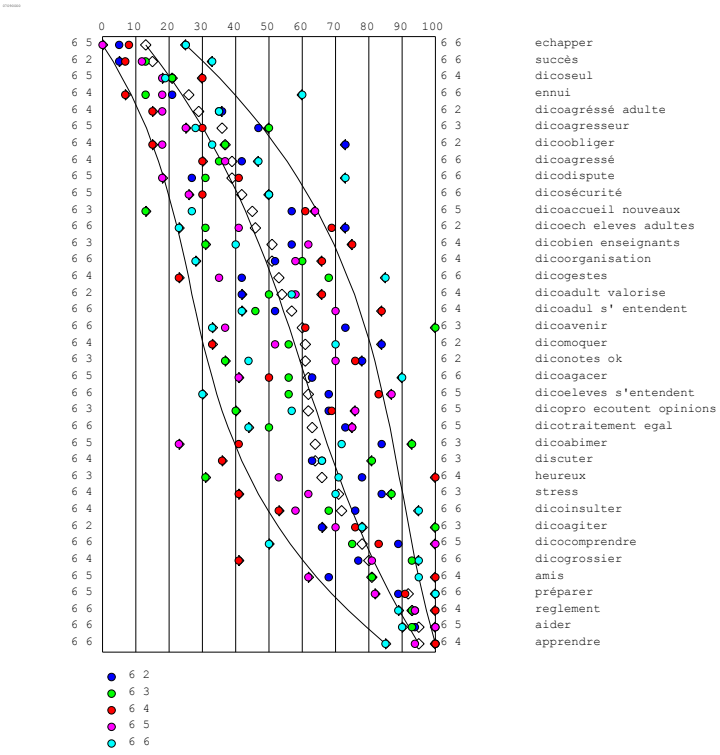
Origine

Il s'agit moins ici d'un outil d'élaboration statistique que d'un instrument de présentation de résultats. Au Printemps 2008, Germain Pauwels nous consultait pour le compte de la société ISAGE, dont l'activité consiste en interventions de recherche et de conseil auprès d'établissements d'enseignement dans les domaines de l'éducation à la santé et de la prévention de la violence et de la toxicomanie.

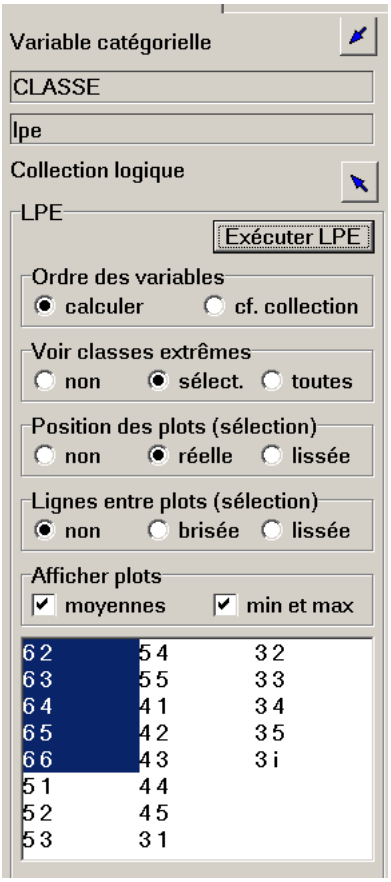
Avec le reste de son équipe, ils avaient mis au point une technique ingénieuse de représentation des attitudes de groupes d'individus (par exemple des classes dans un collège) selon un certain nombre d'indicateurs recueillis par questionnaire, dans la double intention de montrer comment ces indicateurs s'ordonnaient du plus anodin au plus sensible, et comment les différents groupes se situaient au regard de ces indicateurs, permettant de repérer du premier coup d'œil quel groupe était susceptible de poser un problème particulier : une ligne reliait les positions moyennes de l'ensemble des groupes sur les différents indicateurs, et chaque groupe était repéré par un point au dessus, en dessous ou sur la ligne, une position très éloignée de la ligne réclamant une attention particulière. Ils avaient baptisé cette ligne LPE, pour ligne de partage des eaux.

Cette technique était ingénieuse mais extrêmement coûteuse en temps d'élaboration de la représentation, quasiment à la main avec des logiciels évidemment pas conçu pour ça. La demande faite à Hector consistait donc en un outil graphique automatique et interactif, permettant de tester rapidement plusieurs systèmes de paramètres avant d'en choisir un. La partie Consensus a été ajoutée sans avoir été demandée, mais on verra qu'il s'agissait d'un prolongement logique de la démarche. Voici le genre de chose que l'on peut obtenir. On va détailler ci-après comment on y parvient.

```
LPE de lpe selon CLASSE
sous le filtre 6
```


0 10 20 30 40 50 60 70 80 90 100
6 5 6 6 echapper
6 2 6 6 succès
6 5 6 4 dicoseul
6 4 6 6 ennui
6 4 6 2 dicoagréssé adulte
6 5 6 3 dicoagresseur
6 4 6 2 dicoobliger
6 4 6 6 dicoagressé
6 5 6 6 dicodispute
6 5 6 6 dicosécurité
6 3 6 5 dicoaccueil nouveaux
6 6 6 2 dicoech eleves adultes
6 3 6 4 dicobien enseignants
6 6 6 4 dicoorganisation
6 4 6 6 dicogestes
6 2 6 4 dicoadult valorise
6 6 6 4 dicoadul s' entendent
6 6 6 3 dicoavenir
6 4 6 2 dicomoquer
6 3 6 2 diconotes ok
6 5 6 6 dicoagacer
6 6 6 5 dicoeleves s'entendent
6 3 6 5 dicopro ecoutent opinions
6 6 6 5 dicotraitement egal
6 5 6 3 dicoabimer
6 4 6 3 discuter
6 3 6 4 heureux
6 4 6 3 stress
6 4 6 6 dicoinsulter
6 2 6 3 dicoagiter
6 6 6 5 dicocomprendre
6 4 6 6 dicogrossier
6 5 6 4 amis
6 5 6 6 préparer
6 6 6 4 reglement
6 6 6 5 aider
6 6 6 4 apprendre
6 2
6 3
6 4
6 5
6 6

Procédure



Ci-dessus, la partie LPE de l'onglet. La partie Consensus sera expliquée plus loin.

On a besoin d'une variable catégorielle (nominale, ordinale ou logique) que l'on installe en la sélectionnant dans la liste des variables, puis en cliquant la petite flèche bleue appropriée. Dès cette installation, les valeurs possibles de cette variable (étiquettes) s'affichent dans la liste du bas. Celles qui apparaissent ici sur fond bleu ont été *sélectionnées*[8], elles vont pouvoir jouer un rôle spécial.

On a besoin aussi d'une collection de variables logiques (Vrai/Faux), qui peuvent provenir de questions oui/non, ou bien, le plus souvent dans ce type de démarche, de dichotomies sur les médianes[9] de variables numériques issues de valuateurs (cf. note *supra*).

Fondamentalement, la démarche consiste à afficher les variables logiques dans un certain ordre à raison d'une ligne chacune, et de faire figurer sur chaque ligne des éléments généraux comme la moyenne et les *extrema*, ou particuliers comme les positions des différentes classes de la variable catégorielle.

Dès que ceci est en place, on peut cliquer le bouton [Exécuter LPE], et on obtient quelque chose. Ce qu'on obtient dépend des options choisies. Le résultat affiché précédemment correspond aux options telles qu'elles apparaissent ci-dessus, mais de nombreuses autres combinaisons sont possibles.

Ce qu'on obtient toujours, c'est trois lignes noires, qui sont les courbes lissées des positions, pour chaque variable, du minimum, de la moyenne et du maximum. La méthode de lissage est l'ajustement par les moindres carrés à un polynôme du troisième degré.

---

[8] Avec Clic, Shift-Clic ou Ctrl-Clic.

[9] Voir à ce sujet le langage des Formules.

Les options

Ordre des variables

- Calculer : les variables sont ordonnées selon leur fréquence croissante du Vrai
- Cf. collection : les variables sont utilisées dans l'ordre initial de leur collection[10]

Voir classes extrêmes

Les marges de part et d'autre du tableau comportent les étiquettes des classes de la variable catégorielle qui sont titulaires des moyennes minimale et maximale pour cette variable.

- Non : fonction désactivée
- Sélect. : seulement pour les étiquettes sélectionnées dans la liste
- Toutes : les classes extrêmes de chaque variable

Position des plots (sélection)

Si des étiquettes de classes sont sélectionnées, le graphique portera des plots de couleur (avec légende sous le tableau) pour la position moyenne de chacune de ces classes pour chaque variable.

- Non : fonction désactivée
- Réelle : les plots sont à la position réelle moyenne pour chaque classe
- Lissée : les positions sont préalablement lissées selon une cubique

Ligne entre plots (sélection)

Les plots de couleur correspondant à une classe sont reliés par une ligne de même couleur.

- Non : fonction désactivée
- Brisée : une ligne brisée rejoint les plots
- Lissée : la ligne a été lissée

Affiche plots

- Moyenne : selon que la case est cochée ou non, des plots transparents de forme losange marquent la position des moyennes pour chaque variable, pour l'ensemble des classes
- Min et max : selon que la case est cochée ou non, des plots transparents de forme triangulaire marquent les *extrema* pour chaque variable, pour l'ensemble des classes

Sélection

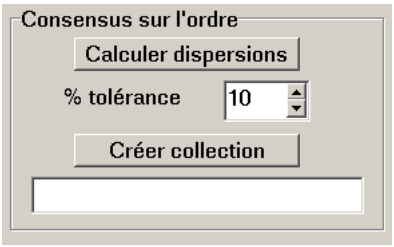
Enfin le fait que certaines étiquettes de classes soient sélectionnées est en soi une option, puisque d'autres option ci-dessus y sont relatives.

De manière générale, comme la plupart des opérations dans Hector, la LPE peut être calculée sous l'influence d'un filtre[11] : c'est le cas ici, où seules les sixièmes sont concernées par l'affichage.

[10] Ordre qui peut être modifié dans la page VARIABLES.

[11] Un filtre est une variable logique posée explicitement en filtre, de sorte que seuls les sujets ayant la valeur Vrai à cette variable sont concernés par les opérations effectuées sous ce filtre. Voir à cet égard le Manuel des Traitements de base. Un filtre peut être aussi complexe qu'on veut, puisqu'on peut le construire avec une formule.

Le Consensus



Sur la même collection logique, et pour les classes actuellement sélectionnées par le filtre s'il en est un, les opérations de consensus posent la question : existe-t-il un ordre des variables pour lequel les sujets seraient à peu près d'accord ? Cette question est importante puisqu'elle peut permettre d'élaborer un ordre canonique permanent (et non pas calculé à chaque fois) des variables, en éliminant celles qui ne font décidément pas consensus.

Le bouton [Calculer dispersions] provoque l'affichage suivant :

```
Analyse du consensus sur l'ordre des variables de lpe selon les positions
de CLASSE
sous le filtre 6
  0,4   3% echapper
  7,0  31% dicogrossier
  2,0   4% succès
  7,8  25% dicoinsulter
…
22,9  26% dicoagréssé adulte
 24,4  24% dicoagresseur
 24,3  30% ennui
 30,1   9% amis
 32,1   2% reglement
 33,2   3% aider
 26,9  35% dicoseul
 34,6   3% préparer
 35,6   4% apprendre
```

On a abrégé le milieu de la liste.

Pour chaque variable, la liste affiche son rang moyen dans l'ordre des fréquences croissantes de Vrai, et le pourcentage qui suit est une traduction de l'écart-type des positions des variables, rapporté à une valeur 100 pour le nombre d'éléments de la liste. La variable « dicoseul », avec une dispersion de plus du tiers de l'étendue, est très peu consensuelle, tandis que « règlement » occupe le rang moyen 32,1 avec très peu de variation : sa position fait donc l'objet d'un consensus relatif.

Créer une liste restreinte à ordre consensuel

Tout dépend de la dispersion qu'on tolère : un dispositif [% tolérance] permet de l'ajuster, avant de cliquer le bouton [Créer collection], à condition d'avoir saisi en dessous un intitulé pour la nouvelle collection, qui regroupera dans l'ordre affiché les variables dont le taux de dispersion n'excède pas la tolérance affichée.

Une telle collection peut servir ensuite de collection de référence pour un usage de la LPE avec une option d'ordre des variables « cf. collection ».

LE VOLET CONNECTIVITE

Origine

Plusieurs besoins simultanés de recherches contemporaines (IUFM, FUNOC), avec un point commun : quand on a recueilli un grand nombre d'indicateurs d'éléments non directement tangibles tels que des compétences ou des attitudes, comment peut-on regrouper au mieux ces indicateurs pour restituer avec la plus grande force possible des variables latentes qui se rapprocheraient de la mesure de ces éléments hypothétiques. Un élément de déclic a été la lecture du mémoire d'habilitation de Sophie Morlaix[12], qui applique aux items des tests CE2 une analyse en piste causales, pour laquelle « une phase préparatoire consiste à étudier individuellement chaque corrélation et à dresser ainsi une cartographie de l'ensemble des situations présentes. Le principe de cette étape préalable est d'identifier des blocs de relations au sein desquelles on retrouve le plus souvent les mêmes items. Cette procédure revêt, de fait, un caractère systématique puisque, pour chaque item, on identifie tous les autres items qui lui sont associés dans les corrélations. Au terme de cette phase, on aboutit à des groupement d'items fortement corrélés entre eux ; certains de ces groupements faisant intervenir un nombre important de corrélations, alors qu'à l'opposé, d'autres ne contiennent que deux ou trois corrélations, celles-ci n'ayant pu être rattachées à d'autres regroupements. ». A la lecture, on se demande quel algorithme a été employé pour analyser cette demi-matrice de 14535 corrélations sur 171 variables ; à vrai dire, ça n'est pas dit clairement, tout se passe comme si ce travail avait été fait à la main, ce qui laisse rêveur ; à la limité, on peut penser qu'on a employé une sorte d'algorithme *agglutinant*, dans lequel un item appartient à un groupe s'il a une corrélation d'un niveau arbitrairement considéré comme suffisant avec au moins un item du groupe. C'est en tous cas ce que suggère la suite immédiate du texte : « Au total, 29 blocs d'items ont été identifiés. A titre d'illustration, le tableau 5 présente l'un de ces blocs d'items (il s'agit de huit items reliés les uns aux autres par neuf corrélations bivariées). ». Il faut entendre, bien sûr, neuf corrélations bivariées acceptables parce qu'égales ou supérieures à un seuil fixé à l'avance. Or, entre huit items, on peut calculer 28 corrélations deux à deux. Dans l'exemple, c'est donc un peu plus d'un tiers des corrélations qui sont acceptables. C'est en ce sens que l'algorithme est agglutinant : les relations s'établissent de proche en proche, on associe C à A et B parce que C est bien corrélé à B, même s'il ne l'est pas à A. La robustesse de tels assemblages pose question, et les fonctionnalités relatives à la connectivité proposées ici visent à fournir au chercheur le moyen d'opérer des regroupements d'items en aveugle sur des bases solides[13].

Par connectivité, on entend donc la possibilité de connecter les items entre eux, pour former des agrégats d'items à la fois cohérents (critère formel) et intelligibles (critère sémantique), tant il est vrai qu'il ne sert à rien d'exhiber des entités d'origine mathématique dont on ne peut débattre, parce qu'on n'est pas en mesure de décrire ce qu'elles représentent.

---

[12] *Identifier et évaluer les compétences dans le système éducatif : quels apports pour la recherche en éducation.* Rapport d'habilitation à diriger les recherches, 2007.

[13] Après vérification auprès de l'auteure, il apparaît que le principe de regroupement s'apparentait plutôt à l'algorithme du chou, fleur, saut maximal (voir ci-après).

Les algorithmes

Trois familles d'algorithmes sont proposées ici. Avec leurs variantes, elles proposent sept manière différentes, et paramétrables d'extraire d'une grande liste d'items un certain nombre de groupes intéressants.

Ces trois familles reçoivent ici les noms poétiques d'algorithme du chou-fleur, d'algorithme de la boule de neige et d'algorithme de l'oignon.

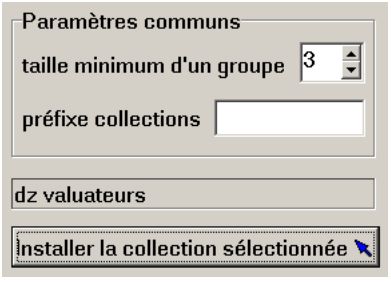
Les algorithmes du chou-fleur consistent d'abord à organiser les items en petits bouquets, puis en bouquets de bouquets de plus en plus gros, jusqu'à la tête de chou fleur, qui est le bouquet suprême. Dans un second temps, on se propose de couper dans le chou-fleur, plus ou moins loin de son cœur, pour séparer les bouquets, qui seront les groupe d'items. Plus on coupe près du cœur, plus les bouquets sont gros et moins ils sont nombreux, et réciproquement. Ce n'est pas propre à l'analyse des items, ça se retrouve aussi dans la construction des typologies[14]. Les différences entre variantes reposent sur différentes manières de considérer à quel point un item est proche d'un bouquet.

Les algorithmes de la boule de neige consistent à partir d'un noyau, qui est toujours constitué des deux variables les plus ressemblantes, et de chercher parmi les items non encore rangés lequel pourrait rejoindre la boule de neige et la rendre encore plus jolie, entendez par là encore plus cohérente, on continue jusqu'à ce qu'aucun item ne mérite de rejoindre la belle boule ; alors on la range sur le côté, et on recommence avec les items en vrac qui restent, jusqu'à ce qu'il n'y ait plus d'items qui se ressemblent assez pour faire une boule.

L'algorithme de l'oignon est un peu l'inverse des précédents : on part de l'ensemble de tous les items, et on épluche les items les moins bien assortis à l'ensemble général, jusqu'à ce qu'on ne puisse plus rien enlever sans nuire à la beauté de l'oignon. Alors on met l'oignon de côté, on ramasse les pluches, on les rassemble et on essaye de faire un nouvel oignon.

Trêve de métaphores. Il est temps d'entrer dans les délices du calcul.

Les paramètres communs



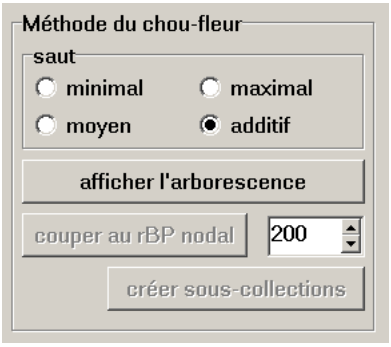
Dans tous les cas, on travaille sur une collection de variables numériques, ou, par extension, de variables calendaires ou logiques (considérées comme numériques aux valeurs 0,1). Il faut donc préalablement en installer une, en la sélectionnant dans la liste des collections puis en cliquant le bouton [Installer la collection sélectionnée].

La taille minimale d'un groupe peut être réglée : c'est le nombre d'items (variables) en dessous duquel le programme s'interdit de constituer un groupe. Cela implique que les variables appartenant à des groupes trop petits, voir des variables isolées, sont exclues de l'analyse. Cela est tout à fait acceptable dans une perspective où l'important est d'avoir des groupes bien construits, et pas spécialement d'exploiter tout le matériau disponible.

[14] Voir le Manuel Factorielles et Typologies. Les typologies sont des analyses de connectivité sur les sujets, et non, comme ici, sur les variables.

Dernier paramètre commun : le préfixe au moyen duquel seront crées les sous-collections représentant les groupes, si on décide de le faire. Ainsi, si on écrit comme préfixe « machin » et qu'il y a quatre groupes, seront créées les collections machin1, machin2, machin3, machin4. Bien sûr on peut créer plusieurs jeux de collections[15] correspondant à divers algorithmes ou paramétrages.

Détail du chou-fleur



L'algorithme de construction du chou-fleur peut être décrit comme ceci :

```
1 Tant qu'il reste des items à examiner dans le tas
      Chercher les deux items lesmieux corrélés,
      les sortir du tas et les associer dans un nœud
      Remettre le noeud dans le tas comme un autre item
      Recalculer les corrélations de tous les autres items restants
      avec l'item issu formé par le nœud récent
      Retourner en 1
```

On constate qu' à chaque itération le nombre d'items simples ou d'items nœud diminue d'une unité, jusqu'à un unique nœud général d'où descendent tous les autres nœuds et items. Dit autrement, un item est un nœud formé d'un seul élément. De plus, la méthode de construction est binaire, et tout nœud qui n'est pas un item simple a exactement deux nœuds descendants.

Le point crucial de cet algorithme est la petite phrase innocente :

```
      Recalculer les corrélations de tous les autres items restants
      avec l'item issu formé par le nœud récent
```

car tout dépend de la méthode de calcul de la corrélation entre nœuds.

Dans tous les cas, si les deux nœuds sont de simples items, la corrélation entre les nœuds est le bon vieux coefficient de corrélation de Bravais-Pearson.

Dans les autres cas, dès qu'au moins un nœud plus complexe qu'un simple item est en jeu, cela dépende la méthode de *saut* choisie :

- Saut minimal : la corrélation d'un nœud existant avec un nœud en création est la plus grande[16] corrélation existante entre le nœud existant et l'un des deux nœuds constitutifs du nouveau nœud.
- Saut maximal : la corrélation d'un nœud existant avec un nœud en création est la plus petite[17] corrélation existante entre le nœud existant et l'un des deux nœuds constitutifs du nouveau nœud.

[15] Une collection n'est qu'une superstructure sur les variables : une variable peut appartenir à plusieurs collections, et la suppression – le démontage – d'une collection n'affecte en rien les variables qui la composent.

[16] Le paradoxe d'appeler saut *minimal* une méthode qui prend la plus grande corrélation est lié au fait qu'il s'agit d'un saut minimal en termes de distance entre objets, et que la corrélation, qui est une ressemblance, est le contraire d'une distance.

[17] Et *vice versa*, bien sûr.

- Saut moyen : la corrélation d'un nœud existant avec un nœud en création est la moyenne (pondérée par le nombre d'items élémentaires sous-jacents) des corrélations existantes entre le nœud existant et les deux nœuds constitutifs du nouveau nœud.
- Saut additif : la corrélation d'un nœud existant avec un nœud en création est la corrélation entre le nœud existant et une variable résultant de l'addition des items constitutifs du nouveau nœud.

Que dire de ces méthodes ? En est-il une meilleure ? Cela dépend de la structure des données qu'on étudie : ainsi le saut minimal est la méthode la plus *agglutinante*, car c'est la moins exigeante pour accepter un item dans un nœud : on entre dans un club si on connaît quelqu'un qui peut vous présenter, même si on ne connaît pas les autres. Elle ressemble fort à ce que l'auteur cité plus haut décrit plus ou moins clairement. Le saut maximal est le plus exigeant, car la corrélation entre deux nœuds est la meilleure des corrélations existantes entre deux items quelconques pris respectivement dans chaque nœud : la perfection ou rien : on entre dans le club seulement si tout le monde vous connaît bien. Le saut moyen et le saut additif font un peu figure d'intermédiaires, le second favorisant de bonnes corrélations item-test dans une analyse de cohérence des groupes.

Mais, à ce stade du propos, c'est-à-dire après avoir cliqué le bouton [afficher l'arborescence], on n'a pas encore de groupes, mais seulement une superstructure générale rassemblant tous les items. Elle s'affiche ainsi, en abrégé[18] :

```
Arborescence des rBP sur la collection dz valuateurs , saut maximal
          dz46 S : je savais déjà ce que je v
        0,125
          dz48 S : je pense que je serai emba
      0,035
          dz20 A:payé ce n'est pas négligeabl
        0,055
            dz67 J.P:ferait mieux de chercher d
          0,201
            dz94 Myriam:emploi ménage enfants
```

les items 67 et 94 forment un nœud, qui fusionne plus haut avec l'item 20, pour former un nœud qui fusionne plus haut avec le nœud formé des items 46 et 48… Chaque nœud porte la valeur de la corrélation entre ses deux descendants immédiats.

On peut analyser cette arborescence à la main, et s'en servir pour découper des groupes : ce sera déjà mieux que d'éplucher la matrice des corrélations. Cependant, on préfère souvent procéder à divers essais de coupure selon diverses méthodes, et juger sémantiquement du résultat : généralement cela ne prend pas très longtemps, car l'expérience montre que chaque ensemble de données à ses caractéristiques propres qui le rendent plus aisé à exploiter avec l'une ou l'autre méthode. Quand on clique enfin le bouton [couper au rBP nodal], éventuellement après avoir ajusté le seuil de coupure, on obtient quelque chose comme ceci :

```
Groupes sur la collection dz valuateurs, saut maximal, seuil 0,300

Groupe 1 : 4 variables ; 0,327
dz34 A : savoir que faire après
dz36 A : aider les enfants pour les
dz31 me débrouiller seule avec les
dz33 A : apprendre à chercher du tr
```

[18] On n'a affiché ici que le début de l'arborescence, qui comporte 199 ligne pour 100 items de départ.

```
Groupe 2 : 7 variables ; 0,462
dz86 voisine :formation inutile
dz89 Maria:à quoi bon à son âge?
dz93 Myriam:pas grave pcq femme
dz37 A : la formation sert pas à gr
dz5 RI:on m'a obligé
dz8 RI:je ne sais pas pourquoi
dz16 RI:mon employeur m'envoie

Groupe 3 : 6 variables ; 0,318
dz21 A:pas payé je ne viendrais pas
dz53 E : salaires trop bas obstacle
dz47 S : la formation ne me servira
dz49 S : plusieurs formations, tour
dz18 O:personne ne sait
dz72 Josette: feux de l'amour
```

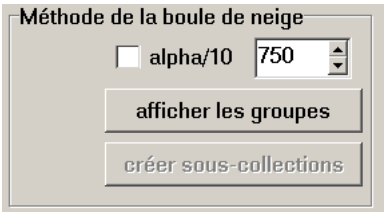
etc. Il y a 9 groupes en tout dans cet exemple.

Chaque groupe porte la corrélation entre les deux branches du nœud principal, calculée selon la méthode qui a été choisie par l'usager.

On peut dès lors générer les collections contenant les variables, en cliquant le bouton [créer sous-collections]. On peut aussi procéder plutôt à d'autres essais. Modifier le seuil de coupure donne moins de groupes, plus resserrés mais plus cohérents. Si on change la méthode de calcul, il faut faire recalculer l'arborescence, qui est différente selon chaque méthode.

Détail de la boule de neige

La boule de neige ne produit pas d'arborescence, mais directement des propositions de groupes.



Deux variantes de l'algorithme sont proposées, selon que la case [alpha10] est cochée ou non. Cependant, la structure fondamentale est la même :

```
1 Tant qu'il reste des items à examiner dans le tas
      Chercher les deux items les mieux corrélés,
      les sortir du tas et les associer dans une boule
      Pour chaque item restant dans le tas,
            Si le joindre à la boule améliorerait
            mieux que les autres l'alpha de la boule, l'y joindre
      Si aucun item n'améliore plus la boule, retourner en 1 essayer une
      autre boule
      Si aucune boule ne se forme plus, arrêter
```

La question centrale est : qu'est-ce que l'alpha ?

Si la case [alpha10] n'est pas cochée, l'alpha n'est autre que le coefficient de fiabilité alpha de Cronbach calculé sur les items constituant la boule. Ce coefficient[19] mesure la probabilité que les différents items, additionnés, contribuent ensemble à l'expression d'une variable latente.

[19] Voir le Manuel des Traitements, rubrique Analyse des Tests.

Si la case [alpha10] est cochée, l'alpha n'est pas l'alpha brut calculé selon la formule de Cronbach, mais une mesure qui en est issue, l'alpha10 ou alpha comparable sur une liste de 10 items[20]. De plus, seuls sont acceptées les boules de neige dont l'alpha10 est au moins égal au seuil fixé à côté du bouton.

Détail de l'oignon

L'oignon ne fournit pas non plus d'arborescence, mais directement des groupes. Son algorithme peut se décrire ainsi :

```
Mettre tous les items dans le tas
1 Enlever tous les items restants sur le tas et les rassembler en un oignon
Tant qu'il est possible d'améliorer l'alpha de l'oignon en enlevant un item
      Enlever l'item qui nuit le plus à l'alpha
      Quand aucun item ne peut plus être enlevé,
      Considérer parmi les différents états intermédiaires de l'oignon qui
      respectent un critère minimal de corrélation item-test celui qui
      donne le meilleur alpha, en faire un groupe, le sortir et reprendre
      le reste dans le tas.
      retourner en 1
Si aucun oignon améliorable n'est formé, arrêter
```

C'est en quelque sorte la réciproque de la boule de neige. Ses résultats ressemblent un peu à ceux du chou-fleur, saut additif, et son implémentation ayant été faite de manière, il est assez lent ; il sera optimisé dans une version ultérieure.

Démarches mixtes

Quelle est la meilleure méthode ? Aucune dans l'absolu, tout dépend de la nature des données. Dans un ensemble à liaison faibles, il sera préférable d'utiliser un algorithme agglutinant pour faire apparaître des formes ; dans un ensemble avec beaucoup de liaisons, on préférera des algorithmes plus exigeants, pour obtenir des structures très robustes, quitte à laisser de côté une partie du matériau initial.

En fait, ce n'est pas au logiciel de décider : il est un auxiliaire, et on peut utiliser plusieurs outils successivement. Avec les données qui ont servi à la mise au point de cette spécialité, les résultats les plus satisfaisants[21] ont été obtenus en trois étapes :

L'algorithme de la boule de neige, version alpha10 avec un seuil à 800 donne quatre groupes, deux d'une vingtaine d'items, deux de 7 et 8 items.

Un nettoyage sémantique enlève par ci par là un ou deux items qui gêneraient la dénomination des groupes

Les deux groupes d'une vingtaine d'items sont réanalysés avec l'algorithme du chou-fleur, saut maximal (très exigeant) : chacun d'eux éclate en deux sous-groupes très cohérents.

On se retrouve enfin avec six groupes de 5 à 10 items, porteurs d'alpha de .796 à .917 et de corrélations item-test minimales de .440 à .680. L'addition des items de chaque groupe pour former de nouvelles variables est ainsi légitimée, et la cohérence sémantique des ensembles qui engendrent ces nouvelles mesures permet de les considérer comme exprimant des variables latentes, dont la distribution dans la population va pouvoir être étudiée et recoupée avec d'autres caractéristiques.

---

[20] A qualité de cohérence égale, l'augmentation des items dans une série augmente sa fiabilité et son alpha de Cronbach. Pour comparer la fiabilité de deux séries de longueur différentes, une formule permet de calculer ce que serait l'alpha de la série, à qualité égale, si elle comportait exactement 10 items.

[21] Les critères de satisfaction sont deux : un ensemble d'items doit être aussi cohérent que possible (corrélation item-test minimale élevée, alpha au delà de .800) et aussi intelligible que possible (ce qui se traduit par la possibilité de lui donner un nom) ; il est parfois nécessaire d'arbitrer des compromis entre ces deux critères.

LE VOLET DISTANCES

Origine

Les techniques employées dans l'onglet Distances proviennent en droite ligne d'Adso2[22], le premier ancêtre d'Hector à avoir été publié. Les précédents avaient été, soit anonymes[23], soit diffusés uniquement avec des polycopiés pour les étudiants[24], soit construits ad hoc pour une seule enquête[25]. Adso2 était le dernier de la lignée à échapper aux interfaces fenêtrées qui se sont imposées par la suite, et ont généré successivement Adso 3, Nestor et l'actuel Hector.

Si le noyau d'Adso2 était un logiciel en ligne de commande assez basique dans ses capacités (tris, croisements et statistiques élémentaires), il était accompagné d'extensions qui l'étaient moins. Outre les analyses factorielles, classificatoires et typologiques, ces extensions manifestaient un intérêt tout particulier pour les distances et dissimilarités, leur analyse et leurs représentations graphiques. Les données qui illustrent le manuel des extensions d'Adso2 proviennent de l'enquête régionale sur l'informatique pédagogique en 1989 et du DEA de Germain Pauwels vers la même époque, mais aussi d'une préoccupation plus ancienne de l'auteur pour les itinéraires et les distances qu'on peut calculer sur ces objets, qui allait déboucher sur *l'analyse des séquences*. On pouvait exploiter et analyser des distances entre profils, entre discours, entre sujets, entre variables.

C'est par ce dernier point que l'analyse des distances revient dans Hector², comme un complément illustratif de l'étude de la connectivité et des regroupements de variables numériques, dans une perspective de mise en évidence des fameuses variables latentes[26].

Principe

S'agissant de variables numériques, la démarche repose ici sur l'analyse de la matrice des corrélations. Sur la base de la corrélation de Bravais-Pearson de valeur r entre deux variables, avec $-1 \leq r \leq 1$, on cherche une mesure de dissimilarité[27] δ qui possède les caractéristiques suivantes :

- $\delta ii = 0$ (identité)
- $\delta ij \geq 0$ (positivité)
- $\delta ij = \delta ji$ (symétrie)

L'identité $\delta ii = 0$ signifie que la dissimilarité entre un objet et lui-même est nulle, puisqu'il est intégralement semblable à lui-même. La positivité $\delta ij \geq 0$ indique qu'une dissimilarité ne saurait

---

[22] Dubus, A. (1992), Méthodes et pratique du traitement statistique en Sciences Humaines avec le logiciel Adso. Lille, Les Trois-Monts.

[23] Logiciels pour la thèse de l'auteur et pour celle de Jacques Hédoux, 1978-1980.

[24] Tel3, support de l'enseignement de l'analyse de données en Sciences de l'Education à Lille 3 du milieu des années 80 au début des années 90.

[25] Le Guillaume, élaboré spécialement pour une enquête nationale du Ministère de la Jeunesse et des Sports sur le devenir des animateurs socio-culturels, 1982.

[26] Variables hypothétiques non-mesurables directement, mais que le modèle théorique propose comme source principale des variations d'autres variables observées, les indicateurs.

[27] La dissimilarité est l'antonyme de la ressemblance, au point que beaucoup de dissimilarités sont construites en comptant entre les objets étudiés les points de ressemblance, puis en soustrayant le résultat obtenu du maximum théoriquement possible.

être négative. Si elle est nulle, elle indique la ressemblance maximale. La symétrie $\delta ij = \delta ji$ indique que la distance entre un objet i et un objet j est la même que dans l'autre sens.

On peut constater que toutes les distances physiques sont des dissimilarités, mais que l'inverse n'est pas vrai : une condition supplémentaire pour qu'une dissimilarité soit une distance est l'inégalité triangulaire :

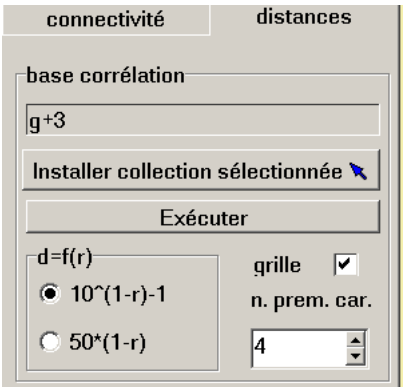
- $\delta ij \leq \delta ik + \delta jk$

Ce qui exprime qu'on peut tracer un triangle dans un plan avec ces trois points, ou encore que c'est toujours plus court et en tous cas pas plus long d'aller tout droit que de faire un détour. C'est évident dans le domaine physique, mais pas forcément en mathématiques.

Or on sait analyser factoriellement un tableau de distances, mais pas un tableau de dissimilarités. On dispose heureusement d'un moyen de corriger un tableau de dissimilarités en distances en ajoutant à chaque distance, sauf à celles qui sont nulles et doivent le rester, la plus petite quantité nécessaire pour que l'inégalité triangulaire soit réalisée pour tous les trios d'objets. Cette transformation, qui conserve l'ordre des dissimilarités, ne pose pas de problèmes particuliers ensuite pour l'interprétation.

L'analyse d'un tableau de distances consiste à projeter l'ensemble des n objets dans un espace à n-1 dimensions respectant intégralement les distances, puis à extraire successivement des dimensions mutuellement orthogonales et d'inertie décroissante, comme dans toute analyse factorielle. On obtient ainsi les coordonnées de chaque objet dans chaque dimension, et on utilise généralement les deux premières (et les plus importantes) comme coordonnées dans un plan illustratif.

Démarche



Le volet Distances étant visible, on sélectionne une collection numérique, et on l'installe en cliquant le bouton approprié.

Avant d'exécuter, on peut choisir le mode de calcul de la distance dans la boîte [d=f(r)] :

Avec l'option 10^(1-r)-1, on obtient :

- Pour r=1, d=0
- Pour r=0, d=9
- Pour r=-1, d=99

Avec l'option 50*(1-r), on obtient :

- Pour r=1, d=0
- Pour r=0, d=50
- Pour r=-1, d=100

Dans les deux cas, 100 (ou 99), représente la distance maximale. ∞ eût été conceptuellement préférable, mais il est techniquement plus difficile à manier.

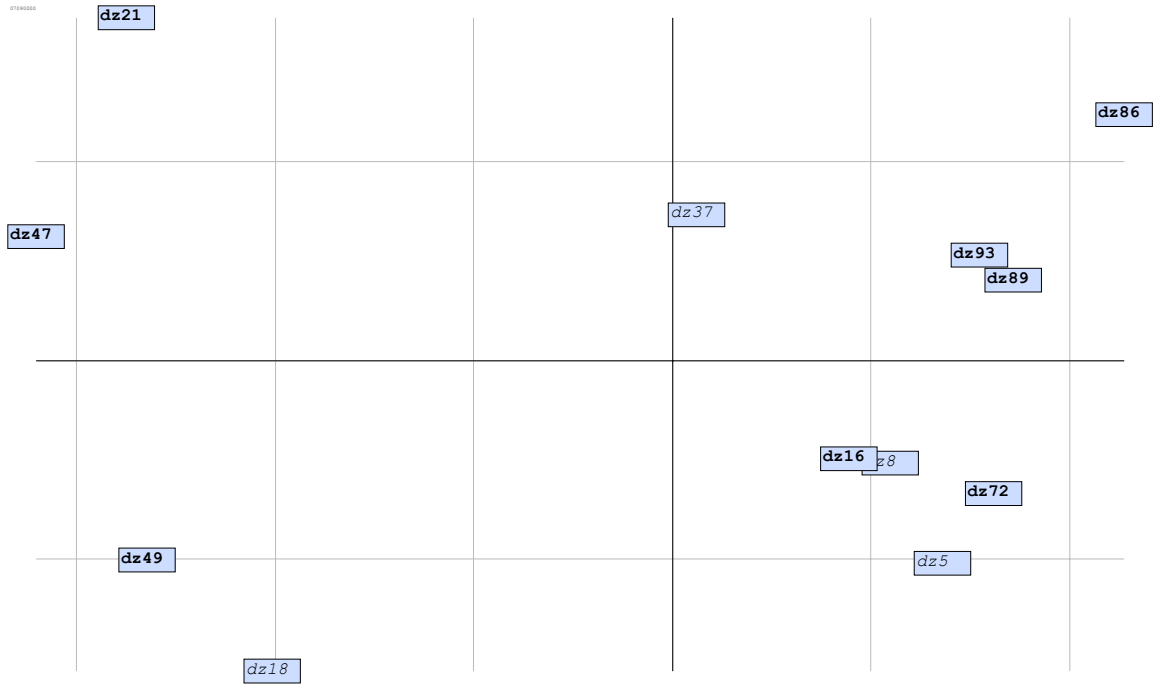
Le choix entre les deux modes de calcul des distances est question d'essais et d'erreurs. Rappelons qu'il n'y a pas de manière parfaite de résumer ces choses en un simple graphique à deux dimensions, et qu'il ne s'agit ici que d'une illustration. La meilleure manière dépendra donc de la nature des données. Selon le cas, il est plus important de mieux séparer les objets très proches, dans d'autres cas c'est le contraire. L'important est d'être au service de l'ergonomie de l'interprétation.

On peut choisir d'afficher ou non la grille des coordonnées. C'est affaire de goût, mais cela facilité généralement la lecture du graphique.

On choisit également le nombre de premiers caractères des intitulés de variable qui seront affichés ; de cela peut dépendre aussi la lisibilité du graphique. Comme on va le voir, il peut être utile de disposer d'intitulés commençant par un numéro, ce qui permet de se référer dans la liste à l'intitulé complet.

A l'exécution, on obtient ce genre de choses :

```
Analyse de distances d = 10^(1-rBP)-1 sur la collection g+3
Correction des inégalités triangulaires = 0,76
   1,6   -0,7     dz72 Josette: feux de l'amour
   0,1    0,7     dz37 A : la formation sert pas à gr
   1,4   -1,0     dz5 RI:on m'a obligé
   1,1   -0,5     dz8 RI:je ne sais pas pourquoi
   0,9   -0,5     dz16 RI:mon employeur m'envoie
   2,3    1,2     dz86 voisine :formation inutile
   1,7    0,4     dz89 Maria:à quoi bon à son âge?
   1,5    0,5     dz93 Myriam:pas grave pcq femme
  -2,0   -1,6     dz18 O:personne ne sait
  -2,8    1,7     dz21 A:pas payé je ne viendrais pas
  -3,2    0,6     dz47 S : la formation ne me servira
  -2,6   -1,0     dz49 S : plusieurs formations, tour
Taux d'inertie conservés par l'axe 1 : 54,72%, par l'axe 2 : 13,50%, par le
plan 68,23%
Seuil de bonne projection : 0,80
```

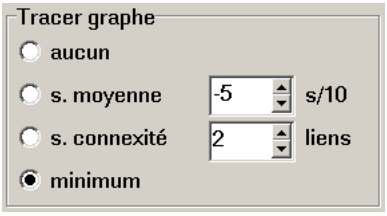


Le texte contient les indications suivantes :

- La correction des inégalités triangulaires a nécessité d'ajouter 0,76 à toutes les dissimilarités non nulles pour en faire des distances.
- La liste des intitulés complets des variables, précédés de leurs coordonnées dans le plan du graphique.
- Les taux d'inertie conservés sur les deux axes et sur le plan qu'ils constituent : c'est une mesure de l'information de distance conservée par l'analyse. Le reste des distances est projetée dans d'autres axes que les deux premiers. De ce fait, la proximité de deux objets dans le plan peut être fallacieuse, et nécessite d'être tempérée par l'usage des bonnes projections.
- Le seuil de bonne projection est la part de l'inertie de la distance d'objet au barycentre du nuage initial qui est conservée dans la représentation en plan. Les objets bien projetés sont écrits en **gras**, ceux qui sont mal projetés le sont en *italique*. Un objet mal projeté n'est pas réellement dans le plan des deux premiers axes : on peut se représenter qu'il est plutôt devant ou derrière, en tous cas dans une autre dimension.

Dans l'exemple ci-dessus, on peut commenter les proximités des objets dz93 et dz89, car ils sont tous deux bien projetés dans le plan. En revanche, la proximité apparente de dz5 et dz72 est sans doute illusoire, car dz5 est mal projeté.

Les graphes



En dessous du panneau précédent, on trouve la commande de sélection des graphes, c'est-à-dire des systèmes de liens reliant les positions des objets. Quatre possibilités s'offrent :

- Aucun graphe
- Le graphe au seuil de la distance moyenne (plus ou moins une certaine part de l'écart-type des distances)
- Le graphe au seuil de connexité, éventuellement avec un certain nombre de liens ôtés
- Le graphe minimal

Quelques définitions[28]

La théorie des Graphes est l'une des branches des mathématiques les plus utiles en informatique, comme en recherche opérationnelle et en statistiques.

Un graphe est une entité qui associe un ensemble de *sommets*, qu'on peut se représenter comme des points, et un ensemble d'*arêtes*, qu'on peut se représenter comme des traits reliant les points.

Un graphe peut être *orienté*, ce qui signifie que les arêtes ont un sens : une arête ij est une arête de i vers j, mais non l'inverse. On représente alors volontiers l'arête par une flèche[29].

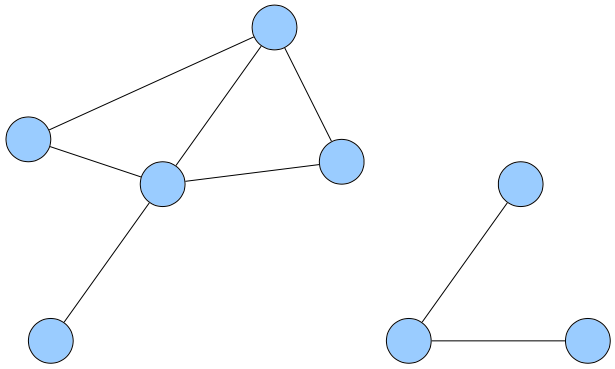
[28] La majeure partie de ce passage est empruntée *verbatim* à Dubus [1992], déjà cité.

[29] Par exemple un graphe représentant le plan de circulation d'une ville, ou les relations amoureuses entre personnages d'une tragédie : certaines arêtes sont, hélas, à sens unique.

Tous les graphes utilisés dans le volet Distances sont non-orientés.

Un *chemin* est une série d'arêtes qui relient deux sommets (généralement *via* d'autres sommets). Un graphe dans lequel existe toujours un chemin entre deux sommets quelconques est *connexe*.

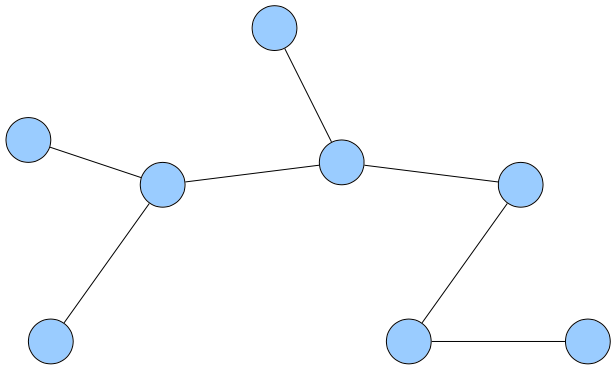
Le graphe ci-dessous n'est pas connexe : chacun de ses morceaux séparés est une composante connexe du graphe.



Un graphe peut être *valué* ou non. Dans un graphe non valué, aucune information n'est fournie sur la longueur des arêtes[30] : elles sont toutes réputées longues de 1 unité, et la *longueur* d'un chemin est simplement le nombre d'arêtes qui le composent. Dans un graphe valué, les arêtes ont chacune une longueur propre, et la longueur d'un chemin est la somme des longueur des arêtes qui le composent.

Les graphes utilisés dans le volet Distances sont valués.

Certains graphes peuvent comporter des cycles, c'est-à-dire permettre de partir d'un sommet et d'y revenir sans passer deux fois par le même sommet, ou, ce qui est équivalent, comporter plusieurs chemins distincts entre certaines paires de sommets. Le graphe ci-dessus comporte des cycles. Un graphe connexe dépourvu de cycles, comme celui qui figure ci-dessous, s'appelle un *arbre*. Si l'un des sommets d'un arbre est explicitement désigné comme la *racine* de l'arbre, ce dernier est dit *planté* en ce sommet.



Un graphe tel qu'entre chaque paire de sommets existe une arête est dit *complet*. *C'est ici qu'intervient le lien avec le problème des dissimilarités et distances.*

Si l'on dispose d'une matrice[31] de dissimilarités mesurées entre tous les éléments d'un ensemble pris deux à deux, rien ne s'oppose à ce qu'on associe à chaque élément un sommet d'un graphe, et, à chaque mesure de dissimilarité sur un couple d'éléments, une arête du même graphe.

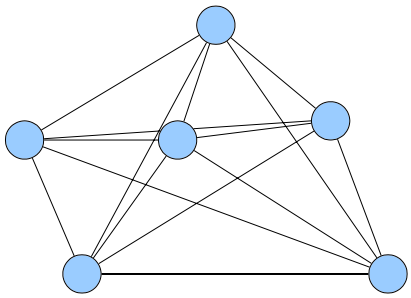
---

[30] Ce peut être le cas d'un graphe décrivant un réseau de relations purement qualitatives, du genre « X peut avoir rencontré Y », très utile par exemple pour étudier la propagation des rumeurs ou la transmission orale des contes.

[31] Il n'est pas nécessaire de considérer la matrice entière, puisque par définition elle est symétrique autour de la diagonale. On travaille donc avec des demi-matrices, mais c'est le problème d'Hector et pas celui du lecteur.

Le graphe des dissimilarités ou distances sera donc valué, puisque les dissimilarités fournissent une longueur à chaque arête. Il y a donc parfaite adéquation entre une demi-matrice de dissimilarités et un graphe complet valué.

Ci-dessous, un graphe complet. On lui a donné peu de sommets, car le nombre d'arêtes, qui est une fonction carrée[32] du nombre de sommets, rend très vite le graphe illisible.



La démarche d'utilisation des graphes de dissimilarités va précisément consister à simplifier le graphe en *supprimant* certaines de ses arêtes, pour mettre en évidence certaines de ses propriétés. Toute la question est de savoir *quelles arêtes* supprimer.

A propos de la lecture des graphes, il est traditionnel de les représenter arbitrairement[33] dans un plan, sans que la proximité des sommets porte la moindre signification s'il ne sont pas reliés par une arête. Un choix différent a été fait pour Hector dans le volet Distances, parce que précisément l'analyse des distances[34] fournit déjà une approximation significative de la position des sommets dans le plan. On projette donc les arêtes du graphe sur le plan construit par l'analyse des distances.

Les seuils de coupure dans le graphe

Un graphe initialement complet dont on a supprimé certaines arêtes est un graphe *partiel*. Si l'on veut conserver l'essentiel de la signification portée par les dissimilarités qui valuent le graphe, il n'est pas question de couper selon la seule commodité : il faut conserver les arêtes correspondant aux dissimilarités qui, d'un point de vue ou d'un autre, sont les plus *importantes*.

Dans une démarche cherchant à identifier quelles variables sont à regrouper, démarche qui est celle du volet Connectivité, ce sont les plus petites dissimilarités qui sont les plus intéressantes. On peut donc fixer une valeur seuil, telle seules les dissimilarités égales ou inférieures soient prises en compte, et leurs arêtes représentées. Le choix du seuil dépend de ce qu'on entend représenter. Hector offre trois variantes de ces graphes à seuils.

Le graphe au seuil de la distance moyenne

On fixe le seuil à la valeur moyenne des dissimilarités : les couples de sommets qui demeurent reliés par une arête peuvent être considérées comme *plutôt proches*, par opposition aux autres couples, considérés comme *plutôt éloignés*. Si les données s'y prêtent, on peut voir apparaître ainsi des sous-groupes homogènes qui s'opposent entre eux.

Si le graphe au seuil de la moyenne demeure trop compliqué, on peut imaginer d'abaisser le seuil, par exemple à la moyenne diminuée d'un écart-type, de manière à ne conserver que les proximités les plus étroites.

Dans l'exemple étudié, le compteur placé à droite du choix [s.moyenne], pour « seuil de la moyenne » porte la valeur –5, qui correspond à –5/10 écarts-types, soit un demi écart-type.

---

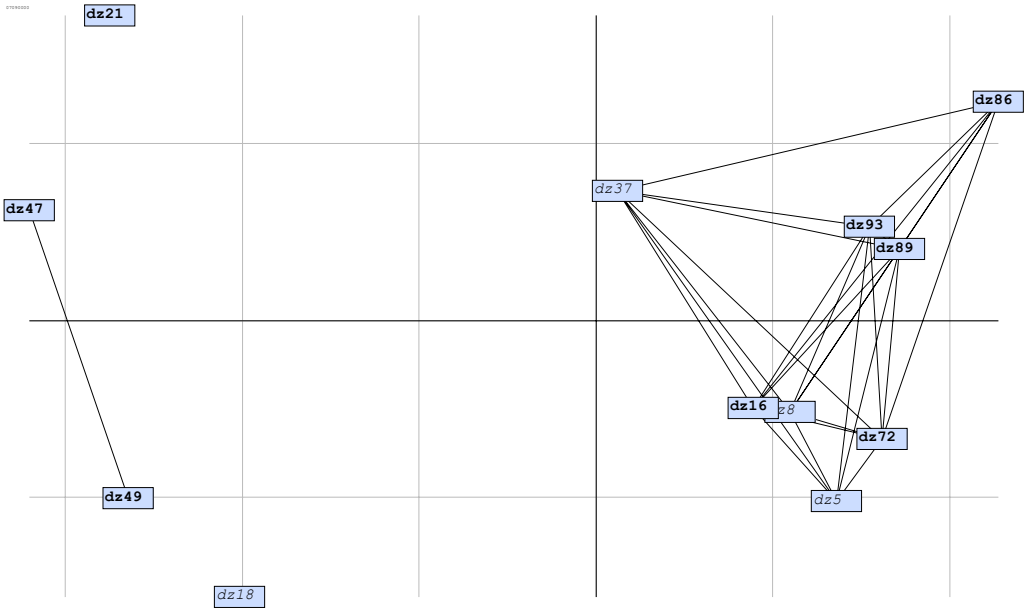
[32] Très exactement $n(n-1)/2$, comme le nombre de poignées de main possibles entre n personnes.

[33] C'était le cas dans les extensions d'Adso2 : les graphes étaient représentés en veillant essentiellement à ne pas emmêler les arêtes.

[34] Ce ne serait donc pas le cas s'il s'agissait d'étudier autre chose que des dissimilarités.

On obtient, sur les mêmes données que précédemment, le graphe suivant :

Graphe au seuil de la distance moyenne -0,5 écart-type

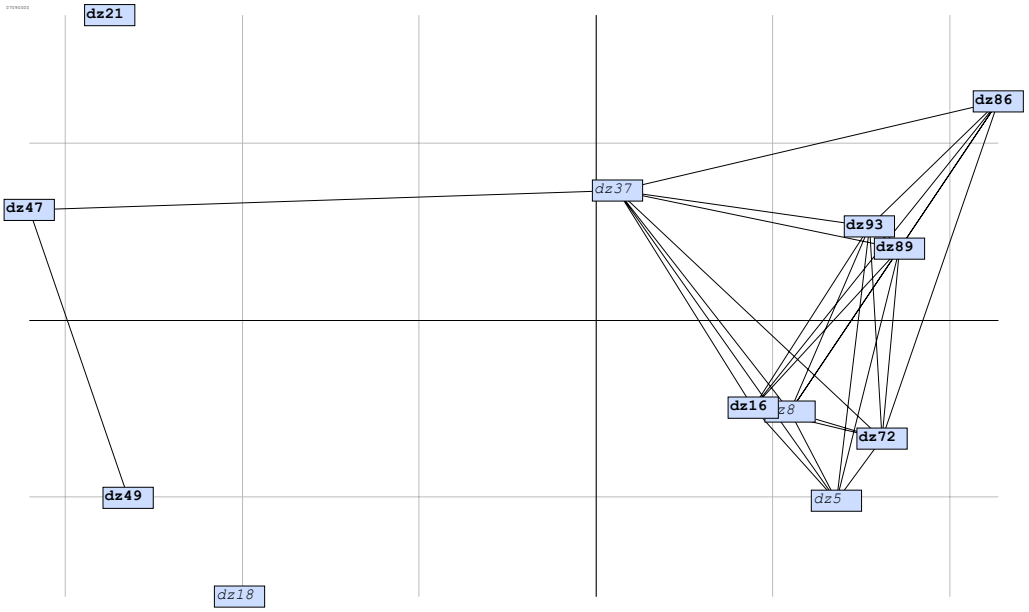


Les données étant ce qu'elles sont, le graphe au seuil de la moyenne n'est pas ici le plus efficace.

Le graphe au seuil de connexité

Au lieu de dépendre d'un critère fixé a priori, le seuil peut aussi être évalué de manière dynamique. C'est le cas du seuil de connexité. Le graphe au seuil de connexité est élaboré de la manière suivante : en partant d'un graphe vide (sans arête), on considère les arêtes par ordre de longueur croissante, et on les ajoute au graphe jusqu'au moment où celui-ci devient connexe, c'est-à-dire où il existe au moins un chemin entre chaque paire de sommets. Comme le graphe obtenu ainsi n'est pas toujours satisfaisant, Hector offre la possibilité d'en ôter les plus longues arêtes, en nombre fixé par le compteur situé à côté du choix [s. connexité]. Le graphe au seuil de connexité constitue donc un point de départ pour une démarche où on enlève quelques arêtes, jusqu'à obtenir un résultat satisfaisant.

Graphe au seuil de connexité, 3 liens coupés



Le résultat n'est pas très différent du précédent, mais les techniques ont très proches.

L'arbre minimum

L'arbre minimum est le graphe connexe sans cycle réunissant tous les sommets initiaux, et tel que la somme des longueurs de ses arêtes soit la plus petite possible. Intuitivement, cela correspond à la représentation la plus économique possible des relations de proche en proche.

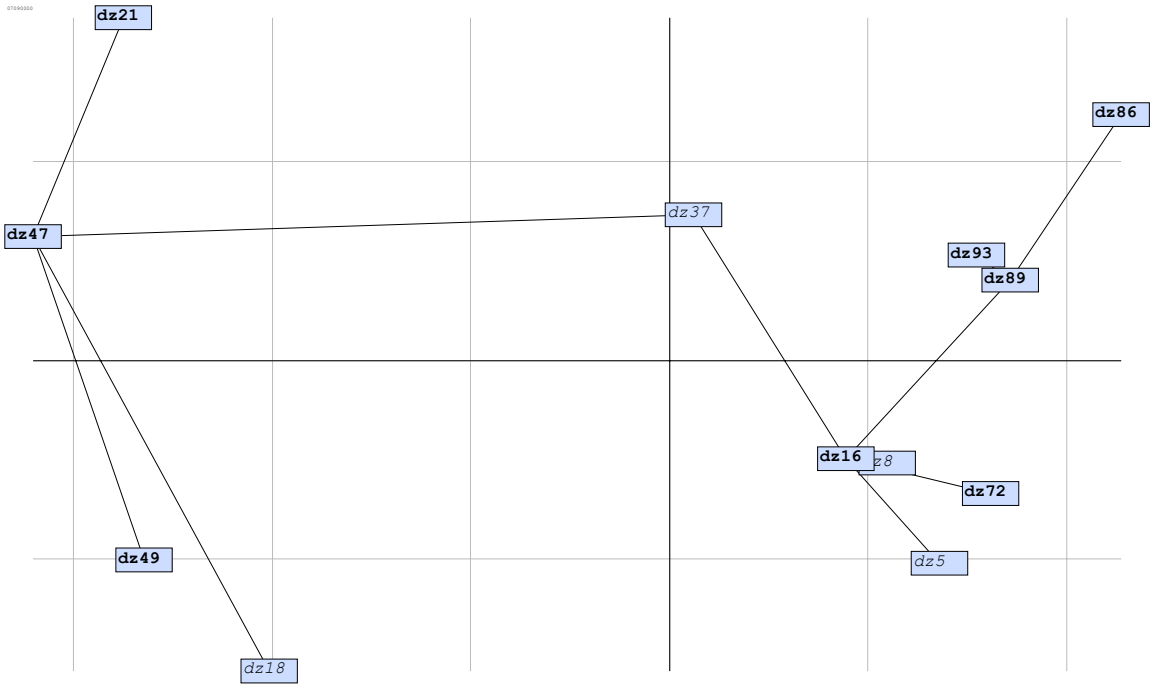
En dépit de cet énoncé un peu savant, la méthode de construction est assez apparentée à la précédente, à ceci près que l'arbre minimum peut comporter certaines arêtes plus longues que d'autres qui ont pourtant été effacées. Kruskal, en 1956, décrit ainsi cette méthode :

- On part d'un graphe vide
- On considère les arêtes dans l'ordre croissant de leur longueur
- S'il existe déjà dans le graphe un chemin entre les sommets qui sont les extrémités de l'arête, on la néglige, sinon on l'ajoute au graphe.
- On arrête dès que tous les sommets sont connectés.

Il s'agit donc en quelque sorte d'un graphe au seuil de connexité sous contrainte d'unicité du chemin entre toute paire de sommets.

Dans le cas des mêmes données que précédemment, on obtient le graphe suivant :

```
Graphe de l'arbre minimal
```



En utilisant le fait que les objets mal projetés sont en italique, on repère l'opposition entre, à gauche, une série 21, 47, 49, avec 18 comme satellite, et, à droite, la série 16, 89, 93, 86, avec 37, 8,5 et 72 comme satellites.

Le fait que l'algorithme accepte des distances plus longues que d'autres qui ont été négligées permet de repérer le groupe de gauche, même s'il est moins homogène que celui de droite.

L'arbre minimum semble le graphe le plus intéressant dans le cas de ces données. Il ne s'ensuit pas qu'il le serait dans tous les cas : cela dépend de la structure propre des données.

Rappelons enfin, encore une fois, que cette technique est illustrative et non démonstrative, et qu'elle vient renforcer une démarche visant à construire des ensembles de variables possédant la double qualité d'être statistiquement cohérents et sémantiquement intelligibles

TABLE DES MATIERES

Le volet Dichotomies 3
- Origine 3
- Principe de la dichotomie 3
- Ce que fait le programme 4
- Les organes et la manoeuvre 5
- Lecture des résultats 6
- Stratégies d'utilisation 7

Le volet Collections parallèles 8
- Origine 8
- Organes 8
- Démarche 8
- Interprétation 11

Le volet Méta-formule 12
- Origine 12
- Démarche 12

Le volet LPE/Consensus 15
- Origine 15
- Procédure 16
- Les options 17
- Le Consensus 18

Le volet Connectivité 19
- Origine 19
- Les algorithmes 20
- Les paramètres communs 20
- Détail du chou-fleur 21
- Détail de la boule de neige 23
- Détail de l'oignon 24
- Démarches mixtes 24

Le volet Distances 25
- Origine 25
- Principe 25
- Démarche 26
- Les graphes 28

Quelques définitions 28
Les seuils de coupure dans le graphe 30

Table des matières 33